## Lab 2.3 – Building and Testing MLB leaning models

In this project, the objective was to predict whether a player would win a Gold Glove based on their fielding statistics, utilizing both a Decision Tree Classifier and a Random Forest Classifier. Initially, the dataset was divided into training and test sets using the Training\_Validation column, and irrelevant columns such as awardID, playerID, yearID, POS, and lgID were eliminated to focus on the essential fielding statistics that would serve as features for the classification models.

For model training, both classifiers were fine-tuned using GridSearchCV to optimize their hyperparameters. The parameter grid for the Decision Tree Classifier included max\_depth, min\_samples\_split, min\_samples\_leaf, and class\_weight, enabling the identification of the most effective configuration for the model. The best-performing Decision Tree achieved a score of approximately 0.9849, with min\_samples\_leaf=1 and min\_samples\_split=5 with a balanced class\_weight and no max\_depth, meaning that each leaf node had at least one sample, and a node needed at least five samples to be considered for splitting.

In comparison, the Random Forest Classifier yielded slightly better performance, achieving a best score of approximately 0.9900. This model was configured with n\_estimators=100, max\_depth=none, class\_weight=balanced, min\_samples\_leaf=1, and min\_samples\_split=5. After applying the models to the test set, the Random Forest Classifier demonstrated high overall accuracy at 99%. However, it faced a significant issue with specificity, scoring 0.105, indicating a tendency towards false positives. While the model showed perfect sensitivity (1.0), it struggled to accurately predict instances where a player would not win a Gold Glove.

To evaluate model performance in more detail, a confusion matrix was used, revealing that the model did great at identifying actual Gold Glove winners but failed to distinguish between players who would not win. The ROC AUC score for the "No Gold Glove" class was approximately 0.950. The specificity challenge persisted despite the class\_weight being set to balanced.

In a separate regression task, the goal was to predict player salaries using batting, pitching, and fielding statistics, applying both a Decision Tree Regressor and a Random Forest Regressor. As in the Gold Glove prediction task, the dataset was split into training and test sets, removing irrelevant columns such as salary, playerID, and yearID. The hyperparameters for tuning included max\_depth, min\_samples\_split, min\_samples\_leaf, and the number of estimators for the Random Forest. The scoring metrics included R² and negative mean squared error (MSE) to evaluate model performance.

In this regression task, the Random Forest Regressor outperformed the Decision Tree Regressor, achieving a best score of approximately 0.283, while the Decision Tree reached only 0.193. The best-performing Random Forest model had 100 estimators, no maximum depth, and a minimum of 5 samples per leaf. The min\_samples\_leaf =1, max\_features =sqrt, and the criterion was squared\_error. On the test set, the best Decision Tree had criterion=squared\_error, no max\_depth, max\_features=sqrt, min\_samples\_leaf'=50, min\_samples\_split=5.

The final model was a Random Forest Regressor with an R² of approximately 0.276, MSE of 26.09 trillion, and mean absolute error (MAE) of 3.29 million. Despite the Random Forest's relative success, the low R² scores highlight that neither model effectively captured the variance in player salaries, likely due to the complexity and wide range of salaries within the dataset.

In conclusion, for the Gold Glove classification task, both models demonstrated high predictive accuracy, but the Random Forest performed better overall. However, the struggle with specificity indicates further tuning of hyperparameters despite a balanced class\_weight. Regarding the salary prediction regression task, the Random Forest Regressor again outperformed the Decision Tree, but the low R² scores suggest that more complex models or additional features might be necessary to improve predictions.